

# SCIENTIFIC REPORTS



OPEN

## Phylogenomic Insight into *Salinispora* (Bacteria, Actinobacteria) Species Designations

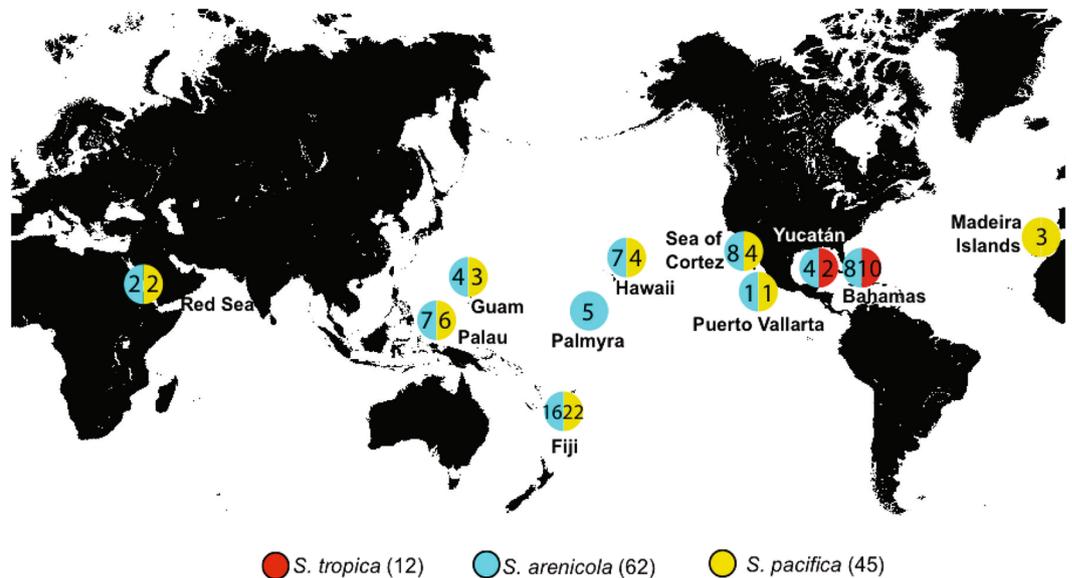
Natalie Millán-Aguiñaga<sup>1,2</sup>, Krystle L. Chavarria<sup>1</sup>, Juan A. Ugalde<sup>1,3</sup>, Anne-Catrin Letzel<sup>1</sup>, Greg W. Rouse<sup>4</sup> & Paul R. Jensen<sup>1,4</sup>

Bacteria represent the most genetically diverse kingdom of life. While great progress has been made in describing this diversity, it remains difficult to identify the phylogenetic and ecological characteristics that delineate groups of bacteria that possess species-like properties. One major challenge associated with species delineations is that not all shared genes have the same evolutionary history, and thus the choice of loci can have a major impact on phylogenetic reconstruction. Sequencing the genomes of large numbers of closely related strains provides new opportunities to distinguish ancestral from acquired alleles and assess the effects of recombination on phylogenetic inference. Here we analyzed the genomes of 119 strains of the marine actinomycete genus *Salinispora*, which is currently comprised of three named species that share 99% 16S rRNA gene sequence identity. While 63% of the core genome showed evidence of recombination, this had no effect on species-level phylogenomic resolution. Recombination did however blur intra-species relationships and biogeographic resolution. The genome-wide average nucleotide identity provided a new perspective on *Salinispora* diversity, revealing as many as seven new species. Patterns of orthologous group distributions reveal a genetic basis to delineation the candidate taxa and insight into the levels of genetic cohesion associated with bacterial species.

The concept that bacteria can be grouped into phylogenetically cohesive clusters with properties that allow them to be regarded as “species” remains controversial<sup>1,2</sup>. It is challenging to determine which clusters represent species level units of diversity and if ecological or evolutionary theory can be invoked to explain the circumstances that led to their formation<sup>3</sup>. As Gevers *et al.* lament<sup>4</sup>, “any effort to produce a robust species definition is hindered by the lack of a solid theoretical basis explaining the effects of biological processes on cohesion within and divergence between species”. Nonetheless, identifying meaningful groups of bacteria and ascribing formal Latinized names remains useful in clinical, environmental, and experimental contexts<sup>5</sup>. In the absence of a robust species concept for bacteria, we are left with a series of metrics used to gauge the relatedness among strains and phylogenetic frameworks within which species level units of diversity are often arbitrarily assigned.

It is widely recognized that bacterial species concepts should consider both genetic diversity and ecology<sup>2,6,7</sup>. Buckley and Roberts stated that, “in moving forward with microbial taxonomy, it is critical to determine whether microorganisms cluster in groups with meaningful commonalities or determine what commonalities may be best used to cluster microorganisms into meaningful groups”<sup>8</sup>. The ecotype model states that bacterial species should fall into well-supported sequence clusters that evolve under cohesive processes and are ecologically distinct and irreversibly separated from each other<sup>6</sup>. A fundamental tenant of this model is that ecologically distinct populations can be recognized as clades in phylogenetic trees and that these clades correspond to fundamental units of diversity or species<sup>2,6</sup>.

<sup>1</sup>Center for Marine Biotechnology and Biomedicine Scripps Institution of Oceanography, University of California San Diego, San Diego, California, United States. <sup>2</sup>Universidad Autónoma de Baja California. Facultad de Ciencias Marinas, Ensenada, Baja California, Mexico. <sup>3</sup>Centro de Bioinformática y Biología Integrativa, Facultad de Ciencias Biológicas, Universidad Andrés Bello, Santiago, Chile. <sup>4</sup>Marine Biology Research Division Scripps Institution of Oceanography, University of California San Diego, San Diego, California, United States. Correspondence and requests for materials should be addressed to P.R.J. (email: [pjensen@ucsd.edu](mailto:pjensen@ucsd.edu))



**Figure 1.** Strain origins. Numbers of strains sequenced at each site for each species with totals in parentheses. Modified with permission from Freel *et al.*<sup>24</sup>, Environ. Microbiol. 14:480–493.

Confounding the common ancestry inferred by phylogenetic reconstruction is homologous recombination. While the efficiency of homologous recombination decreases with increasing genetic distance<sup>9</sup>, it nonetheless occurs between different species<sup>10</sup>. The homologous exchange of genes encoding common housekeeping functions creates challenges for species delineations based on single gene phylogenies and led to the use of techniques such as multi-locus sequence analysis<sup>1</sup>. However, even when multiple loci are considered, an accurate model of vertical inheritance can be difficult to depict due to widespread recombination between species<sup>11,12</sup> including ancestral events that have subsequently become fixed among subclades<sup>13</sup>. While the rates of recombination vary widely among bacteria<sup>14</sup>, it remains largely unknown how this process affects species-level phylogenetic resolution when whole genomes are considered.

Whole-genome sequencing has become an indispensable tool for studying genome evolution, genetic diversity, and bacterial species concepts. It has recently been suggested that genome sequences should be used as a source of taxonomic information<sup>15</sup>. One genome-based metric that is gaining acceptance is the Average Nucleotide Identity (ANI) of the sequences shared between strains. It has been shown that an ANI of 95% corresponds to the 70% DNA:DNA hybridization value traditionally used to delineate bacterial species<sup>16</sup> thus establishing a link to bacterial systematics. Genome sequences also provide unique opportunities to generate highly resolved phylogenies, with automated pipelines to build genomic phylogenies from concatenated protein markers now available<sup>17</sup>. While there is no agreement regarding how many genes it takes to generate a robust phylogenomic evolutionary tree, genome sequences provide comprehensive datasets from which to address evolutionary relationships and predict lateral gene transfer events<sup>18</sup>.

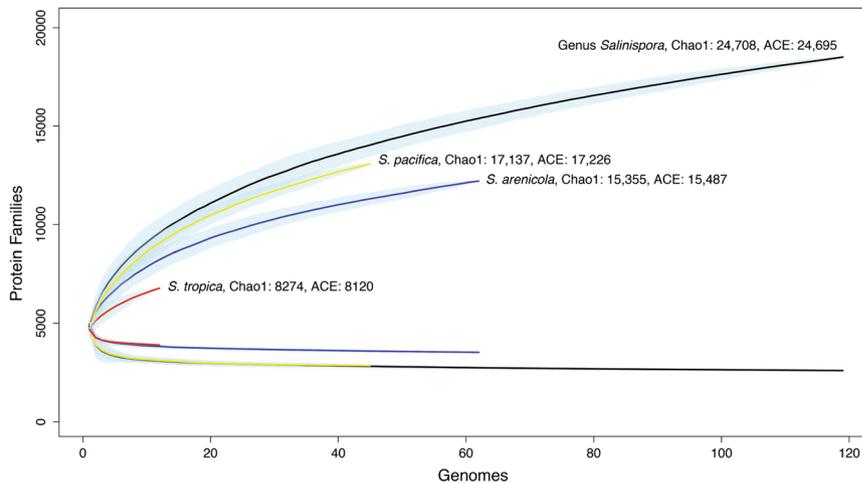
The marine actinomycete genus *Salinispora* provides a valuable model to address bacterial species concepts<sup>19,20</sup>. It is comprised of three closely related species (*S. arenicola*, *S. tropica*, and *S. pacifica*) within the family Micromonosporaceae<sup>21,22</sup> whose relationships could not be confidently resolved based on 16S rRNA gene phylogeny<sup>23,24</sup>. The genus is a rich source of structurally diverse natural products<sup>25</sup>, and there is evidence that certain compounds<sup>26</sup> and their associated biosynthetic gene clusters (BGCs)<sup>27</sup> are fixed at the species level. This has been used to suggest that secondary metabolites represent ecotype-defining traits for *S. tropica* and *S. arenicola*. Similar patterns were not observed for *S. pacifica*<sup>26</sup>, the most diverse of the three species<sup>24</sup>. This greater diversity, coupled with the relatively low recombination to mutation rates observed within the *S. pacifica* clade, were used to suggest it represents an amalgam of ecotypes or newly diverged species<sup>19</sup>. While all three species are prolific in terms of natural product biosynthesis, it was shown that *S. arenicola* differentially invests in interference competition, while *S. tropica* invests in growth thus establishing these co-occurring lineages as distinct ecotypes<sup>28</sup>. Here we present a phylogenomic analysis of the genus *Salinispora* based on the shared gene content among 119 strains. The goals were to assess species level diversity and address the effects of recombination on species level phylogenetic reconstruction.

## Results

**General genome characteristics.** The 119 *Salinispora* genome sequences were derived from 12 *S. tropica*, 45 *S. pacifica*, and 62 *S. arenicola* strains isolated from 11 global locations (Fig. 1). All strains were obtained from marine sediment samples collected at depths from 1–700 meters with the exception of four that were derived from marine sponges (Supplementary Table S1). No heterogeneity was observed in the 2–5 copies of the 16S rRNA gene observed in any of the strains. The draft genome sequences averaged 86.3 contigs (Supplementary Table S2) with the majority of sequence data accounted for by a few large contigs in each genome. The average genome size

Taxa	Genome Size (Mbp)	Gene Count	Scaffold Count	GC Content (%)
<i>Salinispora</i>	5.57	5148	85	69.7
<i>S. arenicola</i>	5.74	5234	80	69.8
<i>S. pacifica</i>	5.42	5079	90	69.9
<i>S. tropica</i>	5.31	4959	89	69.2

**Table 1.** Average genome statistics for the genus *Salinispora* and each species.



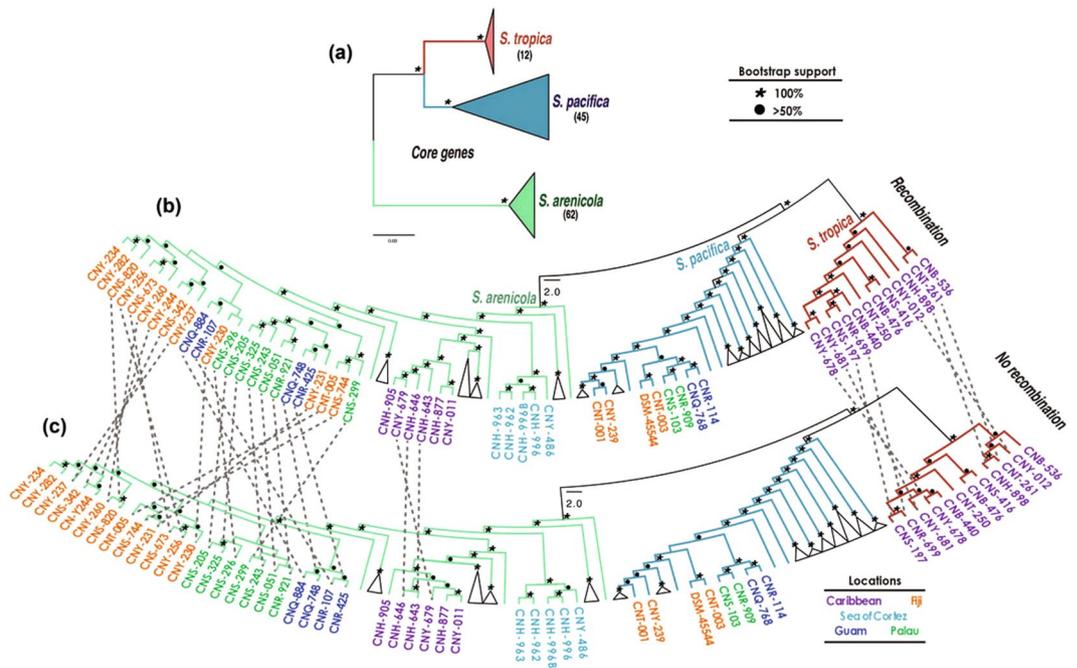
**Figure 2.** Rarefaction curves. Orthologous groups (protein families) plotted vs. the number of sequenced genomes. Core genomes (lower curves) and pan-genomes (upper curves) are shown for the genus and each species. Black: genus, red: *S. tropica*, blue: *S. arenicola*, yellow: *S. pacifica*. Blue shading indicates standard error. Diversity estimates using Chao1 and ACE are given.

was 5.49 Mb, with the *S. arenicola* genomes being larger and containing more genes than the other two species (Table 1).

**Orthologous groups.** The program FastOrtho was used to predict a total of 13,512 orthologous groups (OGs) and 4,980 single copy genes (singletons) among the 119 *Salinispora* genomes revealing a pan-genome that totaled 18,492 protein families. The core genome consists of 2603 OGs shared by all 119 strains, with 2362 of these occurring as a single-copy in each strain. The core genome represents 51% of the average gene content across the genus. Based on the annotation or putative function of the OGs, more than 50% of the pan-genome is comprised of poorly characterized genes (Supplementary Fig. S1). As observed in other genera<sup>29</sup>, the core genome is enriched in functionally annotated genes with the largest group (35%) attributed to metabolism. Similar analyses performed at the species level reveal that *S. tropica* has the largest core genome representing 78.68% of the average gene content while *S. pacifica* displays the smallest core genome at 56.10% of the average gene content while *S. arenicola* was intermediate at 67.42%. As expected, the core genomes vary inversely as a function of the diversity of the strains sequenced within each species.

Rarefaction curves were computed to estimate how effectively gene content had been sampled (Fig. 2). There is clear evidence for saturation when the genus or species-level core genomes are considered and thus the common genetic features that characterize the cultured representatives of these taxa have largely been identified. It is notable that the curves generated from the *S. tropica* and *S. arenicola* core genomes are largely identical, while the curve for *S. pacifica* resembles that describing the genus. For the pan-genomes however, it can be predicted that additional sequencing will reveal additional genetic diversity at all levels. Diversity estimators (Chao1 and ACE) predict more than 24,000 protein families at the genus-level relative to the 18,492 observed. Of the three species, *S. pacifica* shows the highest observed and predicted genetic diversity.

**Effects of recombination on *Salinispora* phylogeny.** The 2362 single copy genes identified in the core *Salinispora* genome (hereafter referred to as the single copy core or SCC) were used to generate a concatenated phylogeny that clearly resolved the genus into three well supported clades in accordance with prior species-level relationships (Fig. 3)<sup>19</sup>. This phylogeny supports the relatively high level of diversity reported for *S. pacifica*. We next used the program PhiPack to address the effects of recombination on phylogenetic reconstruction<sup>30</sup>. This led to the detection of 1,486 SCC genes (62.9%) with evidence of recombination. The remaining 876 genes had no evidence of recombination and are considered the “minimum” core genome. We generated a second concatenated phylogenomic tree using the minimum core genome (Fig. 3) and manually compared this to the individual gene trees for each of the 1,486 SCC genes with evidence of recombination. We identified 635 genes (42.7% of those

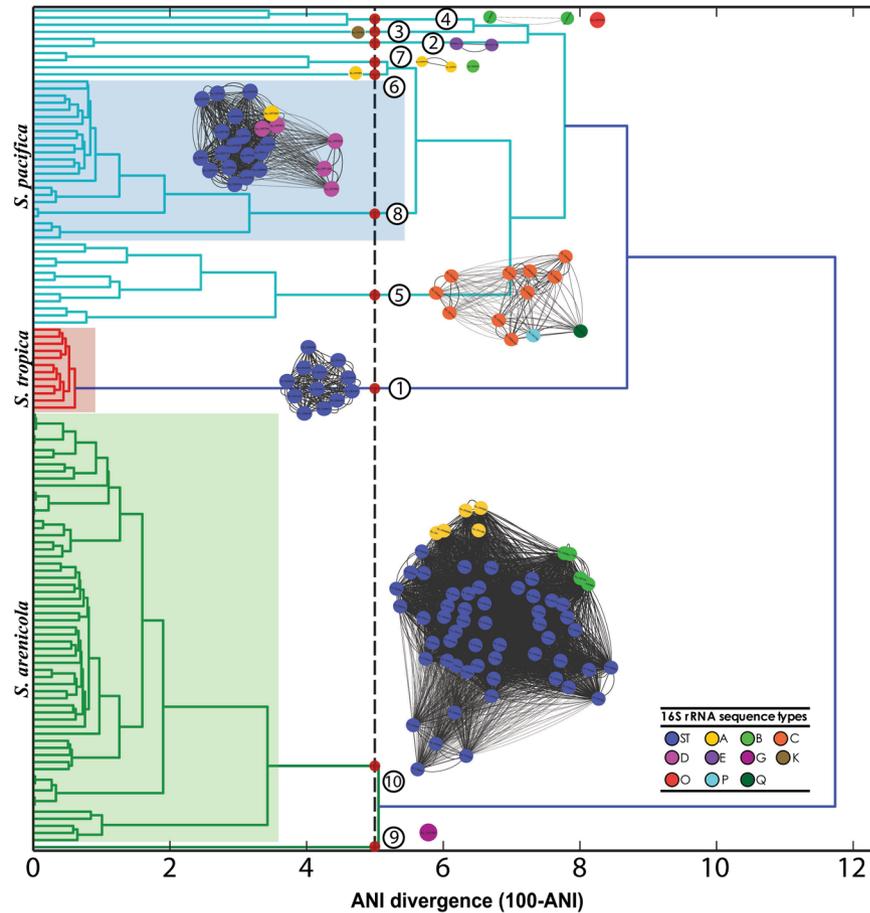


**Figure 3.** *Salinispora* maximum likelihood phylogeny. (a) Collapsed phylogenomic tree based on a concatenation of 2362 shared, single-copy genes. Number of strains analyzed for each species is shown in parentheses. Non-collapsed trees are presented in Fig. 5. (b) Phylogeny based on genes with evidence of recombination. (c) Phylogeny based on genes with no evidence of recombination. Strain numbers are given in cases where the tree topologies differ. When possible, branches with the same topology in both trees were collapsed. Dashed lines depict positional changes of strains in the trees. Branches are color-coded by species. Symbols on the branches represent the support from 1,000 bootstrap replicates. Strain numbers are color-coded by location.

under recombination and 26.9% of the SCC) that displayed incongruent species level phylogenies for at least one strain relative to the concatenated phylogenomic tree (Supplementary Fig. S2). To test for the aggregate effects of recombination, a third concatenated phylogeny was generated using the 1486 SCC genes with evidence of recombination (Fig. 3). Surprisingly, all three trees were both congruent and similarly well supported in terms of the three major clades associated with the named *Salinispora* species. Thus, recombination did not affect *Salinispora* species-level phylogenomic resolution. The large numbers of genes that displayed incongruent species-level phylogenies were insufficient to affect interspecies relationships when taken in the context of the larger gene pools. Notably, the tree generated from the minimum core genome reveals clear biogeographic patterns within *S. arenicola* that were obscured when genes subject to recombination were included (Fig. 3).

These phylogenies were based on the concatenation of various gene sets into a single multiple alignment and the estimation of a single tree from this super-alignment. Given that alternative phylogenetic methods can infer different relationships, the data were re-analyzed using ASTRAL (Accurate Species Tree ALgorithm), a coalescent-based method to summarize individual gene trees into a single species tree<sup>31</sup>. ASTRAL identifies the species tree that agrees with the largest number of individual trees and can be more accurate than maximum likelihood analyses when using a concatenated gene set<sup>32</sup>. Given this, we performed a similar set of analyses using ASTRAL, which resulted in trees that were congruent at the species level with the concatenated trees (Supplementary Figs S3 and S4), thus providing further support for these phylogenetic patterns.

**ANI and ANI-AF metrics.** We next asked if the species assignments inferred from the three primary clades observed in the SCC phylogenomic tree, which have been used to distinguish among the three *Salinispora* species<sup>24</sup>, were in accordance with the proposal that ANI values between members of the same species should be  $\geq 95\%$ <sup>16</sup>. A distance matrix based on ANI values reveals a dendrogram with three primary bifurcations that are congruent with the phylogenomic tree (Fig. 4). However, many strains within the three primary clades fall below the 95% ANI metric, suggesting the existence of additional species-level diversity. More specifically, seven branches within the primary *S. pacifica* lineage could be considered distinct species based on this metric. The most populated branch includes the *S. pacifica* type strain (CNR-114)<sup>22</sup> and 22 additional strains isolated from seven of the global collection sites. The second most populated branch includes 12 strains recovered largely from Fiji while the remaining five branches include one to three strains. The strains comprising these seven lineages are clearly resolved in the expanded phylogenomic tree (Fig. 5) and suggest that the primary clade sister to *S. tropica* is comprised of as many as seven distinct species of which *S. pacifica* is one. Similarly, the *S. arenicola* clade includes two branches that fall below the 95% ANI level. These consist of the single strain CNY-281 and a second branch that contains all of the remaining *S. arenicola* strains including the type strain. Conversely, the *S. tropica*

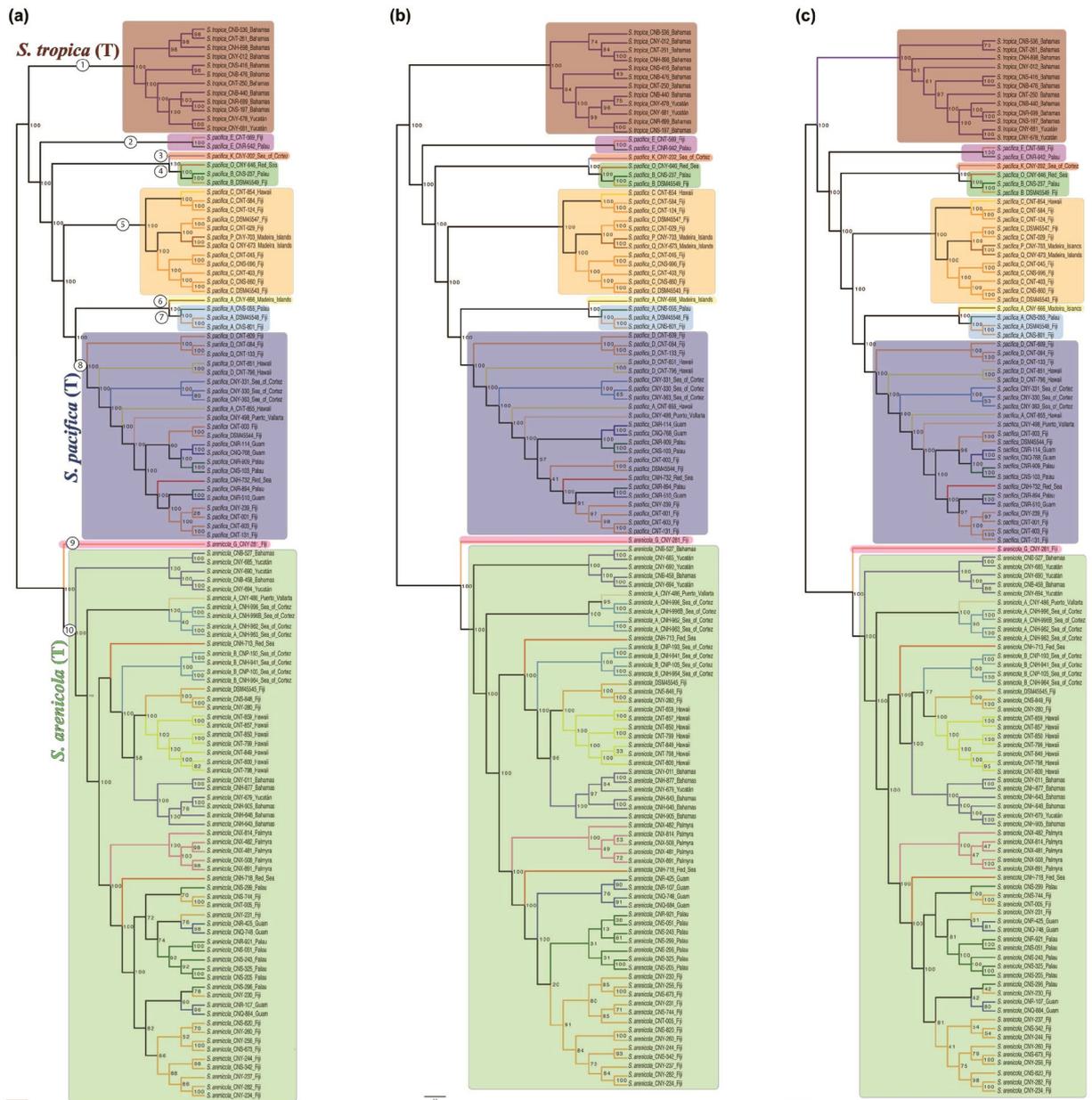


**Figure 4.** Average Nucleotide Identity (ANI) distance matrix. The vertical dashed line represents 95% ANI. Strains associated with the three primary clades are delineated by green (*S. arenicola*), red (*S. tropica*), and light blue (*S. pacifica*) branches. ANI clades that share >95% and are associated with type strains are shaded. Red circles and corresponding numbers represent all lineages that share <95% ANI values including seven (2–7, 9) that do not contain type strains. ANI-AF networks are shown adjacent to the corresponding regions in the dendrogram. Each node represents a strain and is color-coded based on the 16S rRNA gene sequence types (single nucleotide polymorphisms) observed for each species.

clade is represented by a single branch within which all strains share >95% ANI. Thus, according to the ANI analyses, the 119 *Salinispora* strains represent as many as 10 different species.

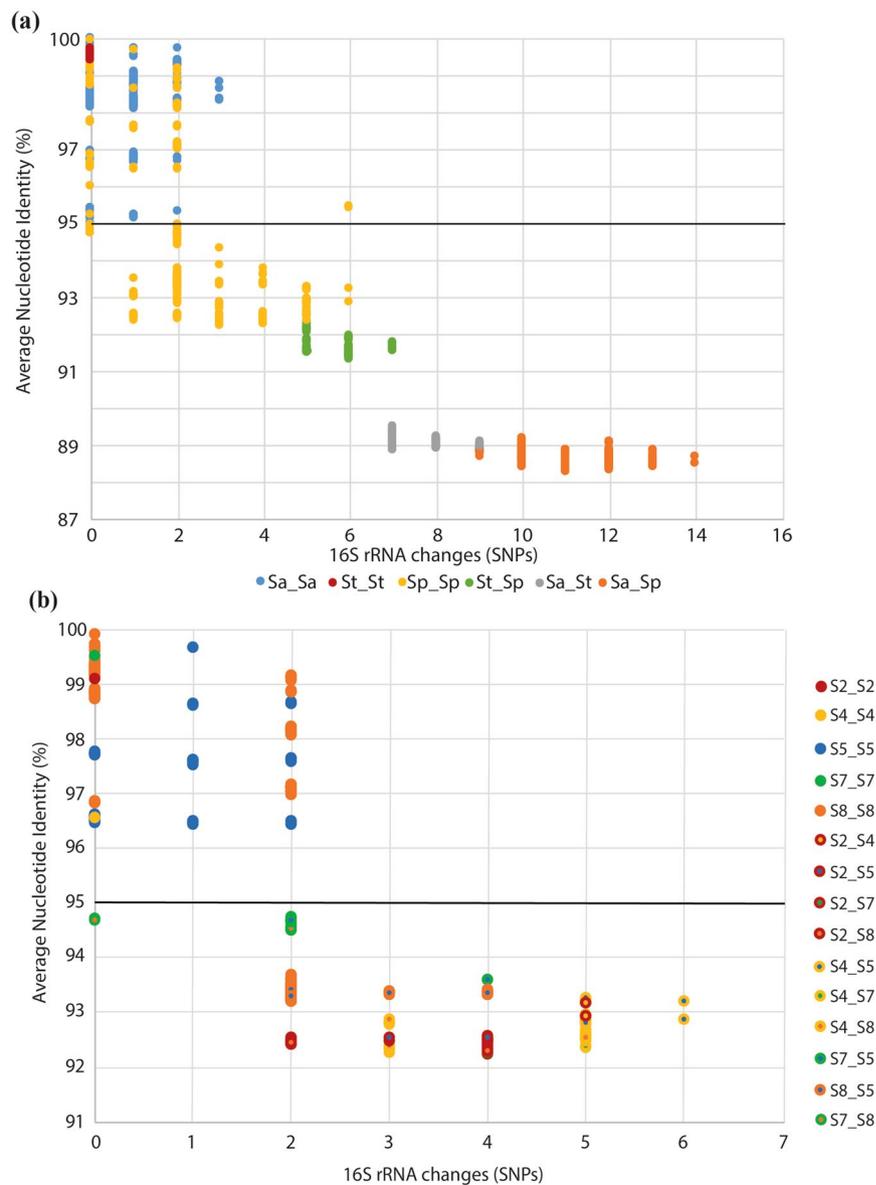
We analyzed the data further using the ANI-AF method<sup>33</sup>, which considers only coding orthologous groups (CDS: From Coding DNA Sequences) and the alignment fraction (AF) between genomes as a measure of relatedness. The values suggested to delineate species are ANI >96.5 and AF >0.6. The ANI-AF results for *S. tropica* and *S. arenicola* remain the same, however within the *S. pacifica* clade, CNS-055 and CNY-646 are delineated as two additional species. Based on the ANI species designations, we re-investigated the effects of recombination on species-level phylogenomic resolution and once again found no effect (Fig. 5). The 10 candidate *Salinispora* species are all clearly resolved both from their minimum core genomes and the SCC genes with evidence of recombination. Thus, recombination does not affect the phylogenetic resolution of the major lineages associated with the three currently named *Salinispora* species or the ten candidate species into which these lineages could be delineated based on ANI.

*Salinispora* 16S rRNA sequence types (single nucleotide polymorphisms) correspond surprisingly well to the ANI-AF clustering (Fig. 4). To further explore these relationships, we plotted 16S rRNA sequence divergence vs. ANI (Fig. 6). Interspecies comparisons based on the three primary clades in the *Salinispora* phylogeny revealed from five (St-Sp) to 14 (Sa-Sp) changes in the 16S rRNA gene. All *S. arenicola* and *S. tropica* intra-species comparisons are above 95% ANI and reveal at most three 16S polymorphisms while many of the *S. pacifica* intraspecies comparisons fall below this line and include up to six SNPs. A linear regression of the data and best-fit line reveals that a 95% ANI value corresponds to 3.1 changes in the 16S rRNA gene (Supplementary Fig. S5). Given that many of the intra-clade comparisons for the major clade that includes *S. pacifica* fall below 95% ANI, we performed a separate analysis of these seven lineages (Fig. 6). As expected, all comparisons within these seven clades fall above and all between-clade comparisons fall below 95% ANI. In this case however, the inter-clade comparisons differ from 0–6 16S rRNA SNPs.



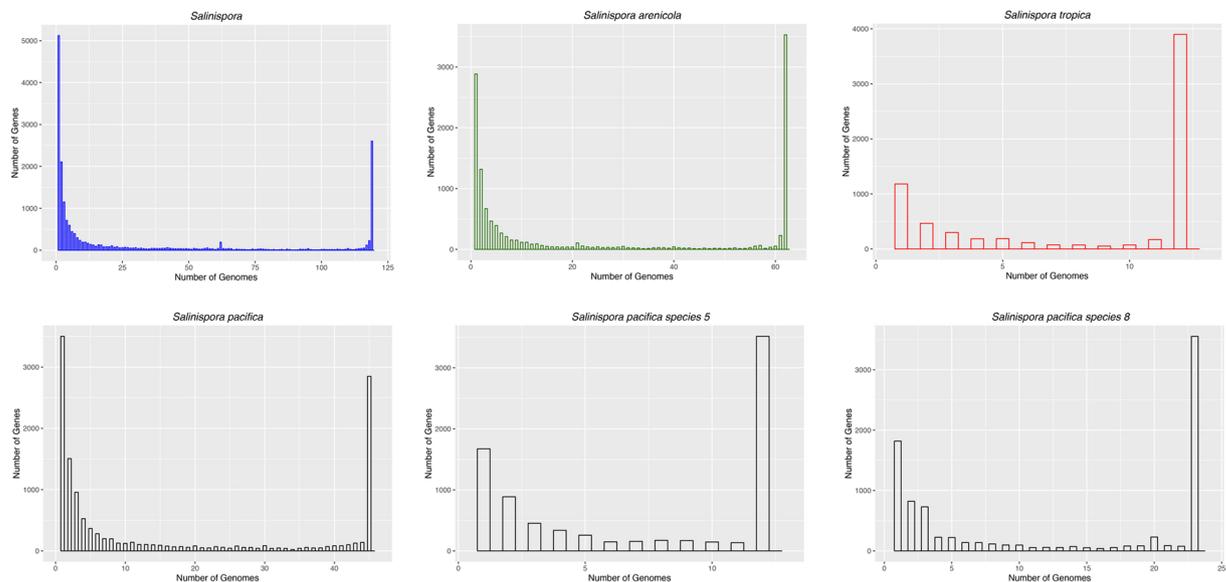
**Figure 5.** Effects of recombination on phylogenetic resolution using ANI species designations. **(a)** Phylogenomic tree based on a concatenation of 2362 shared, single-copy genes. Each sequence or clade that shares <95% ANI with neighboring strains is numbered 1–10 (corresponding to Fig. 4) and shaded with a different color. Species names are listed corresponding to the clades associated with the type strains (T). **(b)** Phylogeny based on genes with no evidence of recombination. **(c)** Phylogeny based on genes with evidence of recombination.

**Genetic basis for species delineations.** We previously reported species-specific patterns of secondary metabolite production in *S. arenicola* and *S. tropica*<sup>26</sup>, however similar patterns were not observed for *S. pacifica*<sup>25</sup>. To further explore this concept in *S. pacifica*, we identified biosynthetic gene clusters (BGCs) associated with secondary metabolism using antiSMASH<sup>34</sup> and manual annotations. We then prepared a similarity matrix using the presence/absence of BGCs in each strain as input (Supplementary Fig. S6). Except for the position of CNY-666, the BGC dendrogram and the phylogenomic tree are largely identical. To further test for evidence of genetic or functional traits that differentiate the candidate *Salinispora* species, we performed similar analyses based on the presence or absence of orthologous groups associated with 23 COG categories (Supplementary Table S3) and found that categories C (energy production and conversion, Supplementary Fig. S7), E (amino acid transport and metabolism), G (carbohydrate transport and metabolism), H (coenzyme transport and metabolism), I (lipid transport and metabolism), and R (general function prediction) consistently delineated the candidate species within the primary *S. pacifica* lineage in accordance with the phylogenomic tree (Fig. 5). Thus, in addition to secondary metabolism, there appear to be major genetic differences among the candidate *S. pacifica* species.



**Figure 6.** Relationships between 16S rRNA diversity and Average Nucleotide Identity (ANI). The black line indicates 95% ANI. **(a)** Inter- and intralade comparisons among the three major lineages represented by *S. tropica* (St), *S. arenicola* (Sa), *S. pacifica* (Sp). **(b)** Inter- and intralade comparisons among the *S. pacifica* clades (2–8) as identified in Figs 4 and 5.

While differences in gene content provide one mechanism to distinguish among species, it can also be expected that the same species will share a certain level of genetic homogeneity. To explore these concepts, we plotted OG distributions across various taxonomic levels (Fig. 7). All histograms clearly show that most genes are either rare or occur in all strains. When the genus is assessed, the core genome represents only 14% of the pan-genome and the relatively large spike in the left portion of the graph provides little evidence for genetic cohesion, as might be expected from a genus comprised of multiple species<sup>29</sup>. Conversely, when *S. arenicola* and *S. tropica* are plotted, the core genomes represent 29% and 58% of the respective pan-genomes, and the numbers of OGs observed in all strains exceed those observed in only one strain. In the primary *S. pacifica* lineage however, the pattern is similar to that detected for the genus, with the core genome representing only 22% of the pan-genome. As was observed in the rarefaction curves, these results are more similar to those for the genus than for either *S. tropica* or *S. arenicola*. We performed similar analyses using the two most populated candidate species within the primary *S. pacifica* lineage and observed OG distributions that resemble *S. tropica* and *S. arenicola*, with core genomes between 40% and 44% of the pan-genomes. These patterns may provide added insight into the levels of genetic cohesion expected for a bacterial species.



**Figure 7.** Numbers of orthologous groups found across all genomes (upper left), *S. arenicola* genomes (upper center), *S. tropica* genomes (upper right), *S. pacifica* genomes (bottom left), *S. pacifica* candidate species “Sp1” (bottom center), *S. pacifica* candidate species “Sp2” (bottom right). The histograms were generated from the pan-genomes excluding singletons and recent paralogs.

## Discussion

The comparison of large numbers of genome sequences derived from closely related bacteria provides a unique opportunity to address bacterial species concepts and the metrics commonly employed to assess sequence-based relationships. Fundamental to this process is the identification of the core genome, which defines the common genomic features that characterize the strains under consideration. As can be expected, core genomes vary widely depending on the diversity of strains and number of genomes examined<sup>35–38</sup>. Nonetheless, this shared gene pool provides unparalleled opportunities to assess levels of sequence divergence and generate comprehensive molecular phylogenies that can be used to infer evolutionary relationships and identify alleles that have been exchanged by homologous recombination.

Homologous recombination provides a mechanism to repair damaged DNA and generate genetic diversity within bacterial genomes<sup>39</sup>. While molecular phylogeny is the primary tool used to assess bacterial diversity, it is well documented that homologous recombination blurs species boundaries and can prevent accurate species delineations<sup>40</sup>. By analyzing the single copy core (SCC) genome associated with 119 closely related *Salinispora* strains, it was possible to generate a detailed and highly supported phylogeny that revealed three primary lineages in agreement with previously established relationships among the three currently named species<sup>19</sup>. Although 63% of the SCC showed evidence of recombination for at least one strain, this had no effect on the evolutionary relationships among the three primary clades. However, removing loci that showed evidence of recombination from the analyses revealed enhanced biogeographic patterning within the *S. arenicola* clade and new evidence for endemism among the structured populations. A majority of genes that displayed evidence of recombination generated phylogenies that were congruent with the established species phylogeny, indicating that most of these events occurred within the three primary lineages as opposed to between them. This is in agreement with the concept that recombination provides a cohesive force that maintains species level units of diversity<sup>41</sup>. However, the large number of core genes that generated incongruent species phylogenies (27%) reveals the importance of selecting the appropriate phylogenetic markers and the power of phylogenomics to overcome this potential source of misleading phylogenetic inference.

ANI analyses revealed that the three primary *Salinispora* clades could be further delineated into as many as 10 different species, all of which could be confidently resolved even when recombinant alleles were included. While three of these lineages are associated with named species<sup>21, 22</sup>, six belong to the relatively diverse clade that is sister to *S. tropica* and contains the *S. pacifica* type strain. This supports the previous suggestion that this clade represents an amalgam of ecotypes or newly diverges species based on its relative low rates of recombination to mutation<sup>19</sup>. The possibility that 10 species are represented among a group of strains that share 99% 16S sequence identity supports the concept that this conserved phylogenetic marker is not the best choice for species-level resolution<sup>4</sup>. Nonetheless, 95% ANI corresponded to approximately three changes in the 16S gene thus indicating that any change in this conserved marker may be meaningful from a taxonomic perspective.

The phylogenetic resolution achieved here is in stark contrast to the genus *Streptomyces*, where high recombination to mutation rates detected using MLSA approaches led to the suggestion that phylogenetic relationships within this genus were better represented by a reticulate network<sup>12</sup>. It remains unclear why the effects of recombination on phylogenetic resolution differ between two taxa within the same bacterial order, however it may relate to the diversity of the strains examined and the number of alleles assessed in the different studies. Furthermore,

it is interesting to speculate that among *Streptomyces* spp., the acquisition of alleles resistant to the many antibiotics they produce may contribute to the high levels of homologous recombination observed, as was shown for the *rpoB* phylogeny in *Salinispora* spp.<sup>19</sup> and exploited to identify the biological targets of secondary metabolites prior to their discovery<sup>42</sup>.

In support of this concept, natural product BGCs are frequently exchanged by horizontal gene transfer<sup>27</sup> and often include a resistant version of the target on which the encoded compounds act<sup>43</sup>. These resistance genes often have homologs in the core genome and can appear as a second copy of a housekeeping gene<sup>44</sup>. In other cases, the resistant housekeeping gene associated with the BGC is the only copy in the genome<sup>19</sup>, suggesting the ancestral allele was subsequently lost. These later events are difficult to distinguish from homologous recombination and may account for some of the single copy genes identified as under recombination in this study. Thus, the ability to produce and be resistant to secondary metabolites may represent a major factor confounding phylogenetic resolution among bacteria enriched in this metabolic capacity. Nonetheless, phylogenomic approaches were sufficient to overcome these incongruences, leading to the generation of stable trees with highly supported clades that can be further evaluated for species-like properties.

Linking strains that can be delineated based on phylogeny or sequence similarity with distinct ecological traits remains a critical and challenging component of microbial ecology. In this regard, it was possible to show that the distributions of secondary metabolite BGCs and six COG categories were largely congruent with the 10 candidate *Salinispora* species delineated based on ANI and resolved in the phylogenomic tree. Thus, there appears to be considerable genetic cohesion among these lineages including within the category of secondary metabolism, which has been reported to represent an important species defining trait for this genus<sup>45</sup>. Ultimately, resolving the genetic and ecological differences among these closely related groups of bacteria, as initially demonstrated between strains of *S. tropica* and *S. arenicola*<sup>28</sup>, will be an essential component of testing the hypothesis that they maintain the properties expected of different species. While it remains to be determined if these results apply more broadly to other groups of bacteria, the expansive growth of genome sequence data will provide ample opportunities to explore species concepts in the future.

## Methods

**Genome sequencing.** Genome sequencing was conducted by the U.S. Department of Energy Joint Genome Institute as part of the Community Science Program (<http://jgi.doe.gov/user-program-info/community-science-program/>). DNA was extracted and the sequence annotation and assembly carried out as previously described<sup>27</sup>. Genomic data is available from the Integrated Microbial Genomes (IMG) database (<https://img.jgi.doe.gov>). IMG genomes ID and NCBI taxon numbers are provided in Supplementary Table S1.

**Orthologous group computation.** A total of 119 *Salinispora* strains (12*S. tropica*, 62*S. arenicola* and 45*S. pacifica*) from 11 different locations (Fig. 1, Supplementary Table S1) were analyzed using the program FastOrtho<sup>46</sup> to identify groups of orthologous protein coding genes (orthologous groups, OGs). This program is a reimplement of OrthoMCL<sup>47</sup> and performs a bidirectional best blast amino-acid analysis. Clustering based on a percent match was performed using default parameters (cutoff = 70, e-value cutoff =  $1e^{-05}$ , and inflation index (I) = 1.5) (<https://github.com/juanu/MicroCompGenomics>). Rarefaction curves and diversity estimates were generated using the vegan package in R (<http://www.R-project.org>). The output matrix of FastOrtho was processed to identify species-specific orthologous groups using an Excel macro (<https://github.com/joseluisrc/FindSharedGenes>). Histograms were plotted from the presence-absence matrix of OGs using the qplot function and the ggplot2 package in R (<http://www.R-project.org>).

**Identification of the core genome and the detection of recombination.** A series of custom python scripts (<https://github.com/juanu/MicroCompGenomics>) were applied to the FastOrtho results to identify the OG members that included gene duplications (paralogs). Orthologous groups that included paralogs were removed to generate the single copy core (SCC) gene pool. The nucleotide sequences of the individual SCC genes in each strain were aligned using MUSCLE with default parameters and trimmed for quality using GBlocks. The SCC genes were screened for evidence of recombination using PhiPack<sup>48</sup>, which included the statistical tests PHI, MaxChi, and Neighbor Similarity Score, all with default parameters. Recombination was inferred when p-values less than 0.01 were detected. Attempts to use the Recombination Detection Program<sup>49</sup> failed due to the large number of loci examined.

**Phylogenetic analyses.** A maximum likelihood (ML) tree was generated for each SCC gene using the program RAxML (command line version) with mid-point rooting and 100 bootstraps (Stamatakis, 2006). The individual gene trees were visualized using the program FigTree v1.3.1 (<http://tree.bio.ed.ac.uk/software/figtree>). Trimmed alignments of each gene were then concatenated and used to build ML phylogenies using RAxML<sup>50</sup> implemented on the CIPRES portal v2.2 at the San Diego Supercomputer Center<sup>51</sup>. Analyses included 1,000 bootstrap replicates using the most complex model (GTR + GAMMA) for both bootstrapping and final ML optimization using default parameter settings. The resulting tree was rooted at the mid-point and visualized using FigTree. Individual SCC gene trees that showed incongruence at the species level with the concatenated tree were scored as under recombination. Two additional concatenated SCC gene trees were then generated for the subsets of this gene pool that included only genes with evidence of recombination and only genes with no evidence of recombination using the methods described above. A similar set of SCC species trees was also generated using the program ASTRAL<sup>31</sup>, which uses the best RAxML trees for each gene tree.

**Average nucleotide identity and alignment fraction.** The average nucleotide identity (ANI) and alignment fraction (AF) were determined for all 119 *Salinispora* genomes using published methods<sup>16,33</sup>. ANI values were calculated for all pairwise comparisons and used to compile a distance matrix representing ANI

divergence (100 - ANI). The custom scripts used to perform these analyses and generate the ANI dendrogram are available ([https://github.com/juanu/ANI\\_analysis/blob/master/ANI\\_blastn.py](https://github.com/juanu/ANI_analysis/blob/master/ANI_blastn.py) and <https://ani.jgi-psf.org/html/download.php>). Cytoscape 3.3.0 was used to visualize the results<sup>52</sup>.

**Clustering based on COG category and functional traits.** The OGs were classified into five major functional categories based on the FastOrtho results and further divided into clusters of orthologous groups (COGs, Supplementary Table S3). These classifications were used to build hierarchical cluster analyses based on the presence/absence of OGs assigned to each COG category using the function `hclust` and the method “average” in the R package (<http://www.R-project.org>). A hierarchical cluster analysis was similarly generated using the presence/absence of secondary metabolite BGCs predicted for the 119 *Salinispora* genomes using antiSMASH<sup>34</sup> as previously described<sup>27</sup>.

## References

- Hanage, W. P., Fraser, C. & Spratt, B. G. Sequences, sequence clusters and bacterial species. *Phil Trans Royal Soc B: Biol Sci* **361**, 1917–1927, doi:10.1098/rstb.2006.1917 (2006).
- Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G. & Hanage, W. P. The bacterial species challenge: Making sense of genetic and ecological diversity. *Science* **323**, 741–746, doi:10.1126/science.1159388 (2009).
- Doolittle, W. F. & Zhaxybayeva, O. On the origin of prokaryotic species. *Genome Res* **19**, 744–756, doi:10.1101/gr.086645.108 (2009).
- Gevers, D. *et al.* Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* **3**, 733–739, doi:10.1038/nrmicro1236 (2005).
- Stackebrandt, E. *et al.* Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Inter J Syst Evol Microbiol* **52**, 1043–1047, doi:10.1099/00207713-52-3-1043 (2002).
- Cohan, F. What are bacterial species? *Annu Rev Microbiol* **56**, 457–487, doi:10.1146/annurev.micro.56.012302.160634 (2002).
- Doolittle, W. F. & Papke, R. T. Genomics and the bacterial species problem. *Genome Biol* **7**, 116, doi:10.1186/gb-2006-7-9-116 (2006).
- Buckley, M. & Roberts, R. Reconciling microbial systematics and genomics. *Amer Acad Microbiol Rep* 2006 (2007).
- Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. & Dowson, C. G. Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J Bacteriol* **182**, 1016–1023, doi:10.1128/JB.182.4.1016-1023.2000 (2000).
- Fraser, C., Hanage, W. P. & Spratt, B. G. Recombination and the nature of bacterial speciation. *Science* **315**, 476–480, doi:10.1126/science.1127573 (2007).
- Doroghazi, J. R. & Buckley, D. H. Widespread homologous recombination within and between *Streptomyces* species. *ISME J* **4**, 1136–1143, doi:10.1038/ismej.2010.45 (2010).
- Cheng, K., Rong, X. & Huang, Y. Widespread interspecies homologous recombination reveals reticulate evolution within the genus *Streptomyces*. *Mol Phylogenetics Evol* **102**, 246–254, doi:10.1016/j.ympev.2016.06.004 (2016).
- Andam, C. P., Choudoir, M. J., Nguyen, A. V., Park, H. S. & Buckley, D. H. Contributions of ancestral inter-species recombination to the genetic diversity of extant *Streptomyces* lineages. *ISME J* **10**, 1731–1741, doi:10.1038/ismej.2015.230 (2016).
- Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**, 199–208, doi:10.1038/ismej.2008.93 (2009).
- Thompson, C. C. *et al.* Microbial taxonomy in the post-genomic era: Rebuilding from scratch? *Arch Microbiol* **197**, 359–370, doi:10.1007/s00203-014-1071-2 (2015).
- Goris, J. *et al.* DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**, 81–91, doi:10.1099/ijs.0.64483-0 (2007).
- Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**, 1 (2008).
- Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* **8**, 163–167, doi:10.1101/gr.8.3.163 (1998).
- Freel, K. C., Millan-Aguinaga, N. & Jensen, P. R. Multilocus sequence typing reveals evidence of homologous recombination linked to antibiotic resistance in the genus *Salinispora*. *Appl Environ Microbiol* **79**, 5997–6005, doi:10.1128/AEM.00880-13 (2013).
- Jensen, P. R. Linking species concepts to natural product discovery in the post-genomic era. *J Ind Microbiol Biotechnol* **37**, 219–224, doi:10.1007/s10295-009-0683-z (2010).
- Maldonado, L. A. *et al.* *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. *Int. J. Syst. Evol. Microbiol.* **55**, 1759–1766, doi:10.1099/ijs.0.63625-0 (2005).
- Ahmed, L. *et al.* *Salinispora pacifica* sp. nov., an actinomycete from marine sediments. *Antonie Van Leeuwenhoek* **103**, 1069–1078, doi:10.1007/s10482-013-9886-4 (2013).
- Vidgen, M. E., Hooper, J. N. A. & Fuerst, J. A. Diversity and distribution of the bioactive actinobacterial genus *Salinispora* from sponges along the Great Barrier Reef. *Antonie Van Leeuwenhoek* **101**, 603–618, doi:10.1007/s10482-011-9676-9 (2012).
- Freel, K. C., Edlund, A. & Jensen, P. R. Microdiversity and evidence for high dispersal rates in the marine actinomycete ‘*Salinispora pacifica*’. *Environ Microbiol* **14**, 480–493, doi:10.1111/j.1462-2920.2011.02641.x (2012).
- Jensen, P. R., Moore, B. S. & Fenical, W. The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. *Nat Prod Rep* **32**, 738–751, doi:10.1039/c4np00167b (2015).
- Jensen, P. R., Williams, P. G., Oh, D. C., Zeigler, L. & Fenical, W. Species-specific secondary metabolite production in marine actinomycetes of the genus *Salinispora*. *Appl Environ Microbiol* **73**, 1146–1152, doi:10.1128/AEM.01891-06 (2007).
- Ziemert, N. *et al.* Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc Natl Acad Sci* **111**, E1130–E1139, doi:10.1073/pnas.1324161111 (2014).
- Patin, N. V., Duncan, K. R., Dorrestein, P. C. & Jensen, P. R. Competitive strategies differentiate closely related species of marine actinobacteria. *ISME J* **10**, 478–490, doi:10.1038/ismej.2015.128 (2015).
- Qin, Q. L. *et al.* Comparative genomics of the marine bacterial genus *Glaciecola* reveals the high degree of genomic diversity and genomic characteristic for cold adaptation. *Environ Microbiol* **16**, 1642–1653, doi:10.1111/emi.2014.16.issue-6 (2014).
- Bruen, T. & Bruen, T. PhiPack: PHI test and other tests of recombination. *McGill University, Montreal, Quebec* (2005).
- Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548, doi:10.1093/bioinformatics/btu462 (2014).
- Kubatko, L. S. & Degnan, J. H. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol* **56**, 17–24, doi:10.1080/10635150601146041 (2007).
- Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences. *Nucl Acids Res*, doi: 10.1093/nar/gkv1657 (2015).
- Weber, T. *et al.* antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucl Acids Res* **43**, 237–243, doi:10.1093/nar/gkv437 (2015).
- Deng, X., Phillippy, A. M., Li, Z., Salzberg, S. L. & Zhang, W. Probing the pan-genome of *Listeria monocytogenes*: new insights into intraspecific niche expansion and genomic diversification. *BMC Genomics* **11**, 1, doi:10.1186/1471-2164-11-500 (2010).
- Donati, C. *et al.* Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol* **11**, 1, doi:10.1186/gb-2010-11-10-r107 (2010).

37. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial Ecol* **60**, 708–720, doi:10.1007/s00248-010-9717-3 (2010).
38. Biller, S. J. *et al.* Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Scientific Data* **1**, 140034, doi:10.1038/sdata.2014.1034 (2014).
39. Rocha, E. P., Cornet, E. & Michel, B. Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet* **1**, e15, doi:10.1371/journal.pgen.0010015 (2005).
40. Hanage, W. P., Fraser, C. & Spratt, B. G. Fuzzy species among recombinogenic bacteria. *BMC Biol.* **3**, 1, doi:10.1186/1741-7007-3-6 (2005).
41. Majewski, J. & Cohan, F. M. DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics* **153**, 1525–1533 (1999).
42. Tang, X. *et al.* Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem Biol* **10**, 2841–2849, doi:10.1021/acschembio.5b00658 (2015).
43. Nett, M., Ikeda, H. & Moore, B. S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat Prod Rep* **26**, 1362–1384, doi:10.1039/b817069j (2009).
44. Kale, A. J., McGlinchey, R. P., Lechner, A. & Moore, B. S. Bacterial self-resistance to the natural proteasome inhibitor salinosporamide A. *ACS Chem Biol* **6**, 1257–1264, doi:10.1021/cb2002544 (2011).
45. Jensen, P. Natural products and the gene cluster revolution. *Trends Microbiol*, doi: 10.1026/j.tim.2106.07.006 (2016).
46. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* gkt1099 (2013).
47. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178–2189, doi:10.1101/gr.1224503 (2003).
48. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681, doi:10.1534/genetics.105.048975 (2006).
49. Martin, D. P. *et al.* RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462–2463, doi:10.1093/bioinformatics/btq467 (2010).
50. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690, doi:10.1093/bioinformatics/btl446 (2006).
51. Miller, M. A., Pfeiffer, W. & Schwartz, T. In *Gateway Computing Environments Workshop (GCE)* 1–8 (IEEE).
52. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504, doi:10.1101/gr.1239303 (2003).

## Acknowledgements

This research was supported by the National Science Foundation (OCE-1235142) and the National Institutes of Health (2U19TW007401 and 5R01GM085770). JU was supported by a Conicyt Grant (Fondecyt Iniciación 11140666) and a research grant from Amazon Web Services. NM-A acknowledges a graduate fellowship from Consejo Nacional de Ciencia y Tecnología (CONACyT-213497). Susana Gaudêncio (REQUIMTE, LAQV) and the Portuguese funding agency FCT/MEC (grant PTDC/QUI-QUI/119116/2010 and IF/00700/2014) are acknowledged for support of sample acquisition from the Madeira Islands, PT. Genome sequencing was conducted by the U.S. Department of Energy Joint Genome Institute and supported by the Office Of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank U. Hentchsel and U. Abdelomohsen for kindly providing strain CNY-646 and W. Aalbersberg and J. Ginigini for strains CNY-230, CNY-234, CNY-256, CNY-282, and CNY-342.

## Author Contributions

N.M.-A. designed and performed the majority of the research, analyzed the data, and drafted the initial manuscript, K.L.C. assisted with generating the genome sequences and the bioinformatic analyses, J.A.U. wrote the scripts and assisted with the bioinformatics analyses, A.-C.L. generated the secondary metabolite gene cluster annotations, G.W.R. assisted with the phylogenetic analyses, P.R.J. helped design the experiments, analyze the data, and write the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at doi:10.1038/s41598-017-02845-3

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017